NAVAL POSTGRADUATE SCHOOL

Monterey, California

THESIS

A SIMULATION STUDY OF ESTIMATES
OF A FIRST PASSAGE TIME DISTRIBUTION
FOR A CENSORED SEMI-MARKOV PROCESS

by

Rick M. Gallagher

September 1986

Thesis Advisor: Patricia A. Jacobs

Approved for public release; distribution is unlimited.

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION: UNCLASSIFIED

1b. RESTRICTIVE MARKINGS

2a SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3 DISTRIBUTION / AVAILABILITY OF REPORT: Approved for public release; distribution is unlimited.

4 PERFORMING ORGANIZATION REPORT NUMBER(S)

5 MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION: Naval Postgraduate School

6b OFFICE SYMBOL *(If applicable)*: Code 55

7a. NAME OF MONITORING ORGANIZATION: Naval Postgraduate School

6c. ADDRESS *(City, State, and ZIP Code)*: Monterey, California 93943-5000

7b. ADDRESS *(City, State, and ZIP Code)*: Monterey, California 93943-5000

8a. NAME OF FUNDING / SPONSORING ORGANIZATION

8b. OFFICE SYMBOL *(If applicable)*

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c. ADDRESS *(City, State, and ZIP Code)*

10 SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO	TASK NO	WORK UNIT ACCESSION NO

11 TITLE *(Include Security Classification)*: A SIMULATION STUDY OF ESTIMATES OF A FIRST PASSAGE TIME DISTRIBUTION FOR A CENSORED SEMI-MARKOV PROCESS

12 PERSONAL AUTHOR(S): Gallagher, Rick M.

13a TYPE OF REPORT: Master's Thesis

13b TIME COVERED: FROM ________ TO ________

14 DATE OF REPORT *(Year, Month, Day)*: 1986, September

15 PAGE COUNT: 41

16 SUPPLEMENTARY NOTATION

17 COSATI CODES

FIELD	GROUP	SUB-GROUP

18 SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*: Simulation, Censored, Semi-Markov

19 ABSTRACT *(Continue on reverse if necessary and identify by block number)*

This thesis reports on a simulation study of parametric and non parametric estimators of a first passage time distribution for a censored semi-Markov process. Four estimators are proposed and compared; Maximum Likelihood Estimator, Renewal Equation Estimator, Asymptotic Renewal Estimator, and the Kaplan-Meier Estimator; the last three estimators are nonparametric. For the particular semi-Markov process studied, the Kaplan-Meier estimator of the first passage times appears to be the best for small times and the Asymptotic Renewal estimator appears to be the best for large times. The Maximum Likelihood estimator is sensitive to incorrect model assumptions. All the estimators are sensitive to censoring.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT: ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS

21 ABSTRACT SECURITY CLASSIFICATION: UNCLASSIFIED

22a NAME OF RESPONSIBLE INDIVIDUAL: Patricia A. Jacobs

22b TELEPHONE *(Include Area Code)*: (408) 646-2258

22c OFFICE SYMBOL: Code 55Jc

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

Approved for public release; distribution is unlimited.

A simulation study of estimates of a first passage time distribution for a censored Semi-Markov process

by

Rick M. Gallagher
Lieutenant, United States Navy
B.S., University of Minnesota, 1979

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 1986

ABSTRACT

This thesis reports on a simulation study of parametric and nonparametric estimators of a first passage time distribution for a censored semi-Markov process. Four estimators are proposed and compared; Maximum Likelihood Estimator, Renewal Equation Estimator, Asymptotic Renewal Estimator, and the Kaplan-Meier Estimator; the last three estimators are nonparametric. For the particular semi-Markov process studied, the Kaplan-Meier estimator of the first passage times appears to be the best for small times and the Asymptotic Renewal estimator appears to be the best for large times. The Maximum Likelihood estimator is sensitive to incorrect model assumptions. All the estimators are sensitive to censoring.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

Finite state space semi-Markov models find application in a variety of areas such as queueing theory, reliability, and clinical trials [Refs. 1,2,3]. The application of these models often centers on the distribution of a *first-passage time* to a state or a set of states representing for example the lifetime of a system or the end of a busy period of a server. Suppose that the *observations* of the path of the semi-Markov process are all that is known about the process.

In a number of these areas, data arise that are censored. This happens frequently, for instance, when fitting lifetime distributions either in medicine or in the field of industrial quality control. In medicine, one might be measuring the amount by which some new drug extends the life of terminally ill patients. A certain number of patients are still alive at the end of the experiment, so we do not know how much their lives have been extended overall, and certain others might have died of unrelated causes or have been removed from treatment prematurely. In quality control one might be measuring the distribution of time-to-failure for a sample of integrated circuit chips under conditions that accelerate aging. Again, many of the chips may not have failed by the end of the trial, while others may have failed at the very beginning due to manufacturing defects unrelated to the mechanisms which cause failures in the long run.

This thesis reports the results of a simulation experiment to compare various parametric and nonparametric estimates of the distribution of a first-passage time for a particular semi-Markov process with censoring. The specific simulation model and estimates considered are given in Chapter 2. Chapter 3 contains the details of the simulation experiment and results. Conclusions from the study are given in Chapter 4.

II. NATURE OF THE PROBLEM

A. PROBLEM

Suppose we observe N individuals. Let $X_t(i)$ be the state of the i^{th} individual at time t. We will assume $\{X_t(i),\ t \geq 0\}$ i = 1, 2, ..., N, are independent identically distributed semi-Markov processes with three states {0,1,2}. The individuals start at $t=0$ in state 1. Upon leaving state 1, they transition to state 0 with probability θ and to state 2 with probability $1-\theta$. From state 2, transition is to state 1 with probability 1. State 0 is an absorbing state. The sojourn time in state i has a distribution function F_i (i = 1,2). The individuals are censored independently. The censoring times are exponentially distributed with mean 1/c. The entire path of transitions and sojourn times are observed until the time of censoring, if any. Let D be the first entrance time to state 0. The problem is to estimate the survival distribution $P\{D > t\}$ with the censored data of the N individuals.

B. ESTIMATORS

Four estimators for $P\{D > t\}$ will be described in this section. The first being the Kaplan-Meier estimate [Ref. 4], and the others are Maximum Likelihood, Renewal Equation, and Asymptotic Renewal estimates from a paper by P. A. Jacobs [Ref. 5].

1. Kaplan-Meier Estimate

One nonparametric estimate for censored data is the product limit estimate. Let $U_1, U_2, \ldots, U_n$ be independent identically distributed random variables with distribution G. Let $V_1, V_2, \ldots, V_n$ be independent identically distributed times to censorship. Let

$$Z_i = \min (U_i, V_i) \qquad \text{(eqn 2.1)}$$

and

$$\delta_i = \begin{cases} 0 \text{ if } U_i \leq V_i \\ 1 \text{ otherwise .} \end{cases} \qquad \text{(eqn 2.2)}$$

Let $Z_{(1)} \le Z_{(2)} \le \ldots \le Z_{(n)}$ be the order statistics of $\{Z_i\}$ and $\delta_{(i)}$ be the corresponding order statistic of $\{\delta_i\}$. The Kaplan-Meier estimate of $\overline{G}(t)$ is

$$\overline{G}(t) = \begin{cases} \prod_{\{i:\, Z_{(i)} \le t\}} C(i)^{1-\delta_{(i)}} & \text{if } t < Z_{(n)} \\ 0 & \text{if } t > Z_{(n)} \ \&\ \delta_{(n)} = 0 \\ \text{Undefined} & \text{if } t > Z_{(n)} \ \&\ \delta_{(n)} = 1 \end{cases} \qquad \text{(eqn 2.3)}$$

where

$$C(i) = (n-i)/(n-i+1) \qquad \text{(eqn 2.4)}$$

[Ref. 4:p. 464] and $\overline{G}(t) = 1-G(t)$. If there isn't any censoring, then the product limit reduces to the binomial estimate for each *t*. This estimate applied to the data of the passage times to state 0 for the N individuals will be referred to as the Kaplan-Meier estimate of the distribution of the first passage time to state 0 and denoted as $\hat{P}_k(t) \equiv \hat{P}_k\{D > t\}$.

2. Maximum Likelihood Estimate

In this subsection, the maximum likelihood estimate will be given for the special case when the sojourn time in state *i* is exponentially distributed with mean $1/\rho_i$ $(i=1,2)$.

Let R_{ij} be the number of transitions from state *i* to *j* for one individual. The log likelihood function for the individual is

$$\mathscr{L} = R_{12}\ln(1-\theta) + R_{10}\ln\theta + R_{21}\ln\rho_2 + (R_{10}+R_{12})\ln\rho_1 - \rho_1 T_1 - \rho_2 T_2 \qquad \text{(eqn 2.5)}$$

where T_i $(i=1,2)$ is the total time spent in state *i* before entrance into state 0 or censoring [Ref. 5:p. 2]. The maximum likelihood estimators are

$$\hat{\theta} = R_{10}/(R_{10}+R_{12}) \qquad \text{(eqn 2.6)}$$

$$\hat{\rho}_1 = (R_{10} + R_{12})/T_1 \qquad \text{(eqn 2.7)}$$

$$\hat{\rho}_2 = R_{21}/T_2 \,. \qquad \text{(eqn 2.8)}$$

The maximum likelihood estimate for the survival distribution is [Ref. 5:p. 5 eqn 1.17]

$$\hat{P}_p\{D > t\} = \{\hat{\theta}\hat{\rho}_1/(\hat{\lambda}_1 - \hat{\lambda}_2)\}\{[(\hat{\lambda}_2 + \hat{\rho}_2)/\hat{\lambda}_2]\exp[t\hat{\lambda}_2] - [(\hat{\lambda}_1 + \hat{\rho}_2)/\hat{\lambda}_1]\exp[t\hat{\lambda}_1]\} \qquad \text{(eqn 2.9)}$$

where $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are the roots of the equation

$$\hat{\theta}\hat{\rho}_1\hat{\rho}_2 + y(\hat{\rho}_1 + \hat{\rho}_2) + y^2 = 0 \,. \qquad \text{(eqn 2.10)}$$

The above estimate will be referred to as the parametric estimate and denoted as $\hat{P}_p(t) \equiv \hat{P}_p\{D > t\}$.

3. Renewal Equation Estimate

The probability $P\{D > t\}$ satisfies the renewal equation

$$P\{D > t\} = \overline{F}_1(t) + (1-\theta)\int_0^t F_1(ds)\overline{F}_2(t-s) + (1-\theta)\int_0^t (F_1 * F_2)(ds)P\{D > t-s\} \qquad \text{(eqn 2.11)}$$

where F_i is the distribution of the sojourn time in state i, $\overline{F}_i(t) = 1 - F_i(t)$, and $F_1 * F_2$ is the convolution of F_1 and F_2.

The solution to the renewal equation 2.11 is

$$P\{D > t\} = g(t) + \int_0^t R(ds)\, g(t-s) \qquad \text{(eqn 2.12)}$$

where

$$g(t) = \overline{F}_1(t) + (1-\theta)\int_0^t F_1(ds)\, \overline{F}_2(t-s) \qquad \text{(eqn 2.13)}$$

and

$$R(t) = \sum_{n=1}^{\infty} (1-\theta)^n (F_1 * F_2)^{n^*}(t) \qquad \text{(eqn 2.14)}$$

where $(F_1 * F_2)^{n^*}(t)$ denotes the n-fold convolution of $(F_1 * F_2)$ with itself at time t.

A nonparametric estimate for $P\{D > t\}$ can be obtained by replacing F_i by its Kaplan-Meier estimate and θ by its maximum likelihood estimate in equation 2.12. If the largest sojourn time in state i is censored then the Kaplan-Meier estimate of F_i is not an honest distribution function $(\hat{F}_i(\infty) < 1)$ since the estimate is undefined past the largest sojourn time. In this case the dishonest distribution estimate is used in all the remaining computations which will give a conservative estimate of the survival distribution.

An approximation to equation 2.12 can be found by using a discrete time approximation to R(t) as follows. Let $\delta > 0$ be a constant and let

$$p_n(\delta) = (1-\theta)\,\{[F_1 * F_2](n\delta) - [F_1 * F_2]((n-1)\delta)\}\ . \qquad \text{(eqn 2.15)}$$

Recursively approximate R(t) as follows

$$R_a(0) = 0 \qquad \text{(eqn 2.16)}$$

$$R_a(\delta) = p_1(\delta)$$

$$R_a(n\delta) = \sum_{k=1}^{n} p_k(\delta) + \sum_{k=1}^{n-1} p_k(\delta)\, R_a((n-k)\delta)\ .$$

An approximation to the solution of equation 2.12 using estimates of F_i and θ is

$$\hat{P}_r\{D > t\} = \hat{g}(t) + \sum_{k=1}^{n(\delta)} \{\hat{R}_a(k\delta) - \hat{R}_a((k-1)\delta)\}\, \hat{g}(t-k\delta) \qquad \text{(eqn 2.17)}$$

where $n(\delta)$ is the largest integer less than t/δ [Ref. 5:p. 9 eqn 2.9]. If the number of individuals N or the time t are large, the estimate of equation 2.17 may require a large number of additions of small non-negative numbers. This estimate will be referred to as the renewal estimate and denoted as $\hat{P}_r(t) \equiv \hat{P}_r\{D > t\}$.

4. Asymptotic Renewal Estimate

Let $\hat{F}_i$ be the Kaplan-Meier estimate of F_i and $\hat{\theta}$ be the maximum likelihood estimate of θ; then define

$$\hat{\varphi}_i(\xi) = \int_0^\infty \exp[s\xi]\, \hat{F}_i(ds) \qquad \text{(eqn 2.18)}$$

where again $\hat{F}_i$ may be a dishonest distribution due to censoring of the last sojourn time in state *i*. The asymptotic renewal estimate of the survival distribution is [Ref. 5:p. 11 eq. 3.11]

$$\hat{P}_a\{D > t\} = \exp[t\hat{\kappa}]\, (\hat{b}/\hat{\mu}) \qquad \text{(eqn 2.19)}$$

where $\hat{\kappa}$ is the solution to the equation

$$(1-\hat{\theta})\, \hat{\varphi}_1(\hat{\kappa})\, \hat{\varphi}_2(\hat{\kappa}) = 1 \qquad \text{(eqn 2.20)}$$

and

$$\hat{\mu} = (1-\hat{\theta}) \int_0^\infty \exp[s\hat{\kappa}]\, s\, (\hat{F}_1 * \hat{F}_2)(ds) \qquad \text{(eqn 2.21)}$$

and

$$\hat{b} = (\hat{\theta}/\hat{\kappa})\, \hat{\varphi}_1(\hat{\kappa})\, . \qquad \text{(eqn 2.22)}$$

The $\hat{\kappa}$ for equation 2.19 was found by numerical search using equations 2.18 and 2.20. The above estimate will be referred to as the asymptotic estimate and denoted as $\hat{P}_a(t) \equiv \hat{P}_a\{D > t\}$.

If $\hat{P}_r\{D > t\}$ were exactly the solution of the equation 2.12 with the Kaplan-Meier estimate of F_i and the maximum likelihood estimate of θ being used then

$$\hat{P}_r\{D > t\}/\hat{P}_a\{D > t\} \sim 1 \qquad \text{(eqn 2.23)}$$

as $t \to \infty$ in the case where the Kaplan-Meier estimates are honest distributions.

III. ANALYSIS OF THE PROBLEM

A. SIMULATION

A Fortran program is written to generate and analyze the data for this problem. All simulations are carried out on an IBM 3033AP computer at the Naval Postgraduate School using the LLRANDOM II random number generating package [Ref. 6]. The data for the simulation experiments are generated as follows: Independent exponential censor times with mean 1/c are generated for each individual. The individual starts in state 1 at $t=0$ and an exponential time with mean $1/\rho_1$ is generated for the sojourn time. A comparison between the sojourn and censor time is done; if the sojourn time is smaller, then the sojourn time is recorded; if the censor time is smaller, the truncated sojourn time and the censored death time are recorded. From state 1, if not censored yet, a uniform random number is compared to theta; if less than theta, the process jumps to state 0 and the uncensored death time is recorded; if greater than theta, the process jumps to state 2 and an exponential sojourn time with mean $1/\rho_2$ is computed. The total time (sojourn times in state 1 plus sojourn times in state 2) is compared to the censored time; with the same actions as listed above. From state 2, the process jumps to state 1 and continues until an uncensored or censored death occurs. The times are recorded and the next individual is started. This continues until all N individuals have been generated. The data in each state is sorted in increasing order for ease of program manipulations. If N is small, it is possible for all the sojourn times in a state to be censored or for all the first passage times to state 0 to be censored which results in $\hat{P}_r(t)$, $\hat{P}_a(t)$, or $\hat{P}_k(t)$ being undefined for all t. In these cases the replication is dropped and a new replication generated.

A sample data set is listed below for N = 10. The first row under state 1 and state 2 gives each particular censored or uncensored sojourn time that is generated for that state. Under each sojourn time, the binary number indicates whether the individual is censored (1) or not (0) during that sojourn time. State 0 indicates times of death (passage time to state 0), and whether censored (1) or not (0); note that the times indicate either the time of death (not censored) or the time of censoring (censored death time). The sojourn and death times listed below have been sorted, along with its associated censor indicator.

State 2

0.1629	0.2041	2.2201
0	0	0

State 1

0.1356	0.1615	0.2114	0.2748	0.2996	0.3067	0.3450	0.3725	0.3996	0.4305
1	1	1	0	0	0	1	0	0	1
0.8676	1.1980	2.4630							
0	0	0							

State 0 N = 10

0.1356	0.1615	0.2748	0.3450	0.8676	0.8832	0.9930	1.1980	2.4630	2.7312
1	1	0	1	0	0	1	0	0	1

Using equations 2.3, 2.9, 2.17, and 2.19, estimates of the survival distribution $P\{D > t\}$ from the data are calculated from subroutines in the Fortran program. Output from the program produces a table like the one below that includes: time, actual survival probability (ACT(t)), parametric estimate ($\hat{P}_p\{D > t\}$), renewal estimate ($\hat{P}_r\{D > t\}$), asymptotic estimate ($\hat{P}_a\{D > t\}$), and the Kaplan-Meier estimate of the first passage time to state 0 ($\hat{P}_k\{D > t\}$). The actual survival probability ACT(t) is computed using equations 2.9 and 2.10 with the actual parameter values instead of the estimated values. The Kaplan-Meier estimate uses only the uncensored first passage times to state 0. Output in Table I is for the data set listed above.

In Table I, the renewal and asymptotic estimates decrease as t increases. In this case, the largest sojourn times in both state 1 and state 2 are uncensored. To demonstrate what can happen when the largest sojourn times are censored, Table II shows a case where the largest sojourn times in state 1 and state 2 are censored. Notice that after $t = 5$ there is little change in the renewal estimate. The survival probability levels off and becomes constant. The asymptotic estimate starts low (half the probability) and goes to zero just after $t = 5$. In a third case, when either of the largest sojourn times in state 1 or state 2 are censored, the effects are somewhere

between the two cases mentioned above; the renewal estimate starts to level off but may not become constant and the asymptotic estimate starts lower than normal and may go to zero. The dishonest Kaplan-Meier estimate of F_i has a definite affect on $\hat{P}_r(t)$ and $\hat{P}_a(t)$ for large t.

TABLE I

OUTPUT FROM PROGRAM

Survival Probability $P\{D > t\}$

Time	ACT(t)	$\hat{P}_p\{D > t\}$	$\hat{P}_r\{D > t\}$	$\hat{P}_a\{D > t\}$	$\hat{P}_k\{D > t\}$
.5	0.79965	0.73641	0.66667	0.68473	0.87500
1.0	0.66340	0.56522	0.52606	0.54348	0.58333
2.0	0.47996	0.35318	0.36158	0.34238	0.38889
5.0	0.19737	0.09549	0.09181	0.08560	Undefined
7.0	0.10985	0.04027	0.03451	0.03397	Undefined
10.0	0.04563	0.01103	0.00874	0.00849	Undefined
12.5	0.02194	0.00375	0.00272	0.00268	Undefined
15.0	0.01055	0.00128	0.00086	0.00084	Undefined

TABLE II

OUTPUT FROM PROGRAM

Survival Probability $P\{D > t\}$ (largest sojourn censored)

Time	ACT(t)	$\hat{P}_p\{D > t\}$	$\hat{P}_r\{D > t\}$	$\hat{P}_a\{D > t\}$	$\hat{P}_k\{D > t\}$
.5	0.79965	0.79130	0.78783	0.35320	0.90000
1.0	0.66340	0.65115	0.65547	0.16873	0.78750
2.0	0.47996	0.46358	0.52706	0.03851	Undefined
5.0	0.19737	0.18049	0.49071	0.00046	Undefined
7.0	0.10985	0.09675	0.49020	0.00002	Undefined
10.0	0.04563	0.03797	0.49017	0.00000	Undefined
12.5	0.02194	0.01742	0.49017	0.00000	Undefined
15.0	0.01055	0.00799	0.49017	0.00000	Undefined

B. ANALYSIS

For the simulated model described above, parameter values of $\rho_1 = 1$, $\rho_2 = 1$, $\theta = 0.5$, and $c = 0.5$ are used. The simulation uses two different numbers of observed individuals. The number of individuals is set at 10 and 50, representing a low and moderate number of subjects. The simulation is replicated 500 times utilizing different seeds to generate the data. The average relative bias for each estimate is computed by

$$ARB(t) = (1/M) \sum_{i=1}^{M} (EST_i(t) - ACT(t))/ACT(t) \qquad \text{(eqn 3.1)}$$

where $EST_i(t)$ is the value of an estimate computed for the i^{th} replication at time t and ACT(t) is the actual model value at time t. For the Kaplan-Meier estimate, M is taken as the number of Kaplan-Meier estimates of the first passage time to state 0 still defined by time t. For the other estimates, M is the number of replications (500).

The figures below show histograms of the relative bias of the observations $(EST_i(t)-ACT(t))/ACT(t)$. Figure 3.1a shows histograms of the relative bias for each of the four estimates when N = 10 and at $t = 0.5$. Each of the histograms looks relatively normal with possibly a slight skew to the left. The parametric estimate has the tightest distribution and the asymptotic estimate the worst which is expected since the asymptotic properties are for large t. Figure 3.1b shows the relative bias for each estimate when N = 10 and $t = 5.0$. The parametric is somewhat normal but skewed to the right. The renewal estimate looks a little less skewed. The asymptotic estimate is skewed to the right and looks exponential. At time $t = 5.0$, less than half of the Kaplan-Meier estimate of the first passage time to state 0 are defined. The histogram of the defined Kaplan-Meier estimate is starting to show an accumulation of mass at -1.0 which is the value of the relative bias where the largest passage time observation is uncensored and less than 5.0.

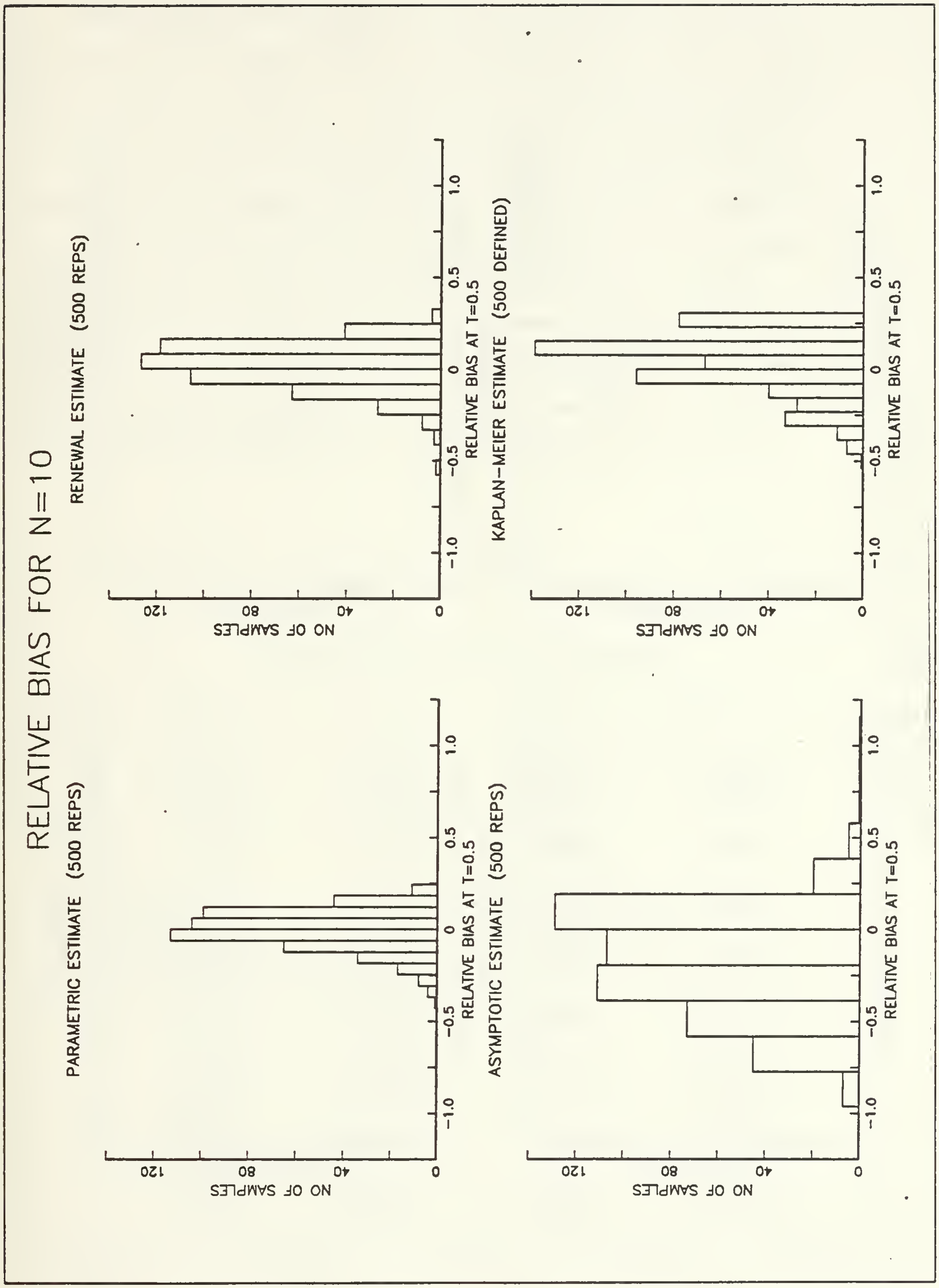

Figure 3.1a Histograms of relative bias for N = 10 and t = 0.5.

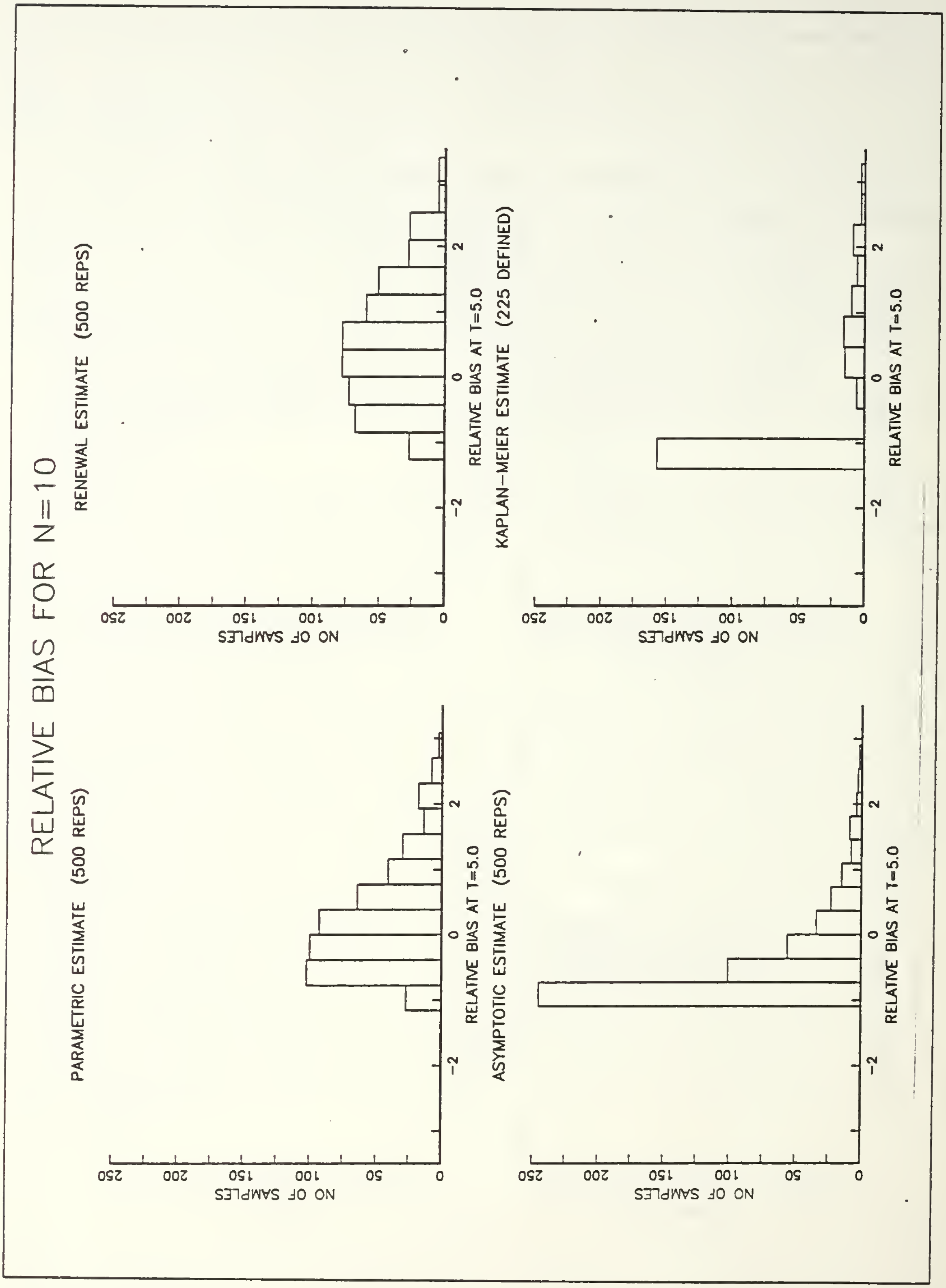

Figure 3.1b Histograms of relative bias for N = 10 and t = 5.0.

Table III shows the ARB(t) of the estimates for the case when N = 10 individuals and Table IV for the case when N = 50 individuals. The ARB(t) for each estimate is given for selected values of *t*. Along with the ARB(t) in the parentheses is the corresponding standard error. The standard error is computed by taking each observation of the relative bias (EST_i(t)-ACT(t))/ACT(t) and subtracting the ARB(t), squaring this and summing over all M observations, then dividing by M-1. This produces the distribution variance, which is divided by M and the square root taken of to get the standard error of the ARB(t) for each estimate at time *t*. The variance together with the average relative bias can be used to obtain an estimate of the relative mean squared error of the estimate. The right most column of the Tables III and IV gives the number of replications out of 500 that still has defined Kaplan-Meier estimates of the distribution of the first passage time to state 0 by time *t*.

TABLE III
AVERAGE RELATIVE BIAS

Exponential Model N = 10 (500 Reps)

Time	$\hat{P}_p(t)$	$\hat{P}_r(t)$	$\hat{P}_a(t)$	$\hat{P}_k(t)$	#KM
.5	-.00183 (.00533)	.01292 (.00545)	-.18069 (.01287)	.01448 (.00743)	500
1.0	.00575 (.00942)	.02788 (.00917)	-.24628 (.01501)	.04135 (.01153)	499
2.0	.03470 (.01656)	.05672 (.01568)	-.34206 (.01942)	.07631 (.01847)	462
5.0	.24324 (.04126)	.52162 (.04267)	-.43114 (.03196)	-.37581 (.07156)	225
7.0	.50476 (.06596)	1.30581 (.07867)	-.39861 (.04399)	-.69076 (.08469)	185
10.0	1.15728 (.12758)	3.84602 (.19504)	-.25101 (.07358)	-1.00000 (.00000)	174
12.5	2.06816 (.21894)	8.49575 (.41214)	-.02144 (.11647)	-1.00000 (.00000)	174
15.0	3.52449 (.37592)	18.05457 (.86673)	.34887 (.18841)	-1.00000 (.00000)	174

The parametric estimate $\hat{P}_p(t)$ uses the most correct information about the process. For $N = 10$, the parametric estimate is within three standard deviations of zero bias for $t < 5$. As t gets larger, the relative bias tends to increase. The parametric estimate understandably has the smallest relative bias for small t. For large t, the small sample sizes involved are probably responsible for the larger relative bias. For small times the renewal estimate and the Kaplan-Meier estimate for the distribution of the first passage time to state 0 have about the same average relative bias. For small N and large t, the renewal estimate has large bias. As noted before, the renewal estimate will be biased if the largest observations of the sojourn times in a state are censored thus causing the Kaplan-Meier estimate $\hat{F}_i$ to be undefined. The bias could also be caused by the step size in the discrete time approximation (step size 0.01) being too large, or by numerical error in summing large quantities of small numbers, as mentioned earlier. The Kaplan-Meier estimate does well for small t and small N. As time increases, the number of data points depreciates rapidly. Because of the small number of subjects in each run, the Kaplan-Meier estimate of the distribution of the first passage time to state 0 lost over half its data due to undefined distributions. By time $t = 10$, there are no survivors using the Kaplan-Meier estimate, resulting in the -1.0 average relative bias. From equation 2.23, the renewal estimate and the asymptotic estimate should be approximately the same for large t if the Kaplan-Meier estimates $\hat{F}_i$ are always defined. The asymptotic estimate is negatively biased for small t but changed over at $t > 12.5$. Once again, it could be biased due to censoring of the largest sojourn times. The asymptotic estimate has the smallest average relative bias for large time t.

Figure 3.2a shows histograms of the relative bias for each of the four estimates when $N = 50$ and at $t = 0.5$. Each of the histograms again looks relatively normal.The distributions are much tighter when compared to Figure 3.1a. The parametric estimate has the tightest distribution and again the asymptotic estimate the worst. Figure 3.2b shows the relative bias for each estimate when $N = 50$ and $t = 5.0$. All the estimates except the Kaplan-Meier estimate look relatively normal with possibly a slight right skew. The Kaplan-Meier estimate of the first passage time to state 0 has just over two thirds of its distributions defined and is showing the start of an accumulation at -1.0 due to the largest passage time to state 0 being less than 5.0.

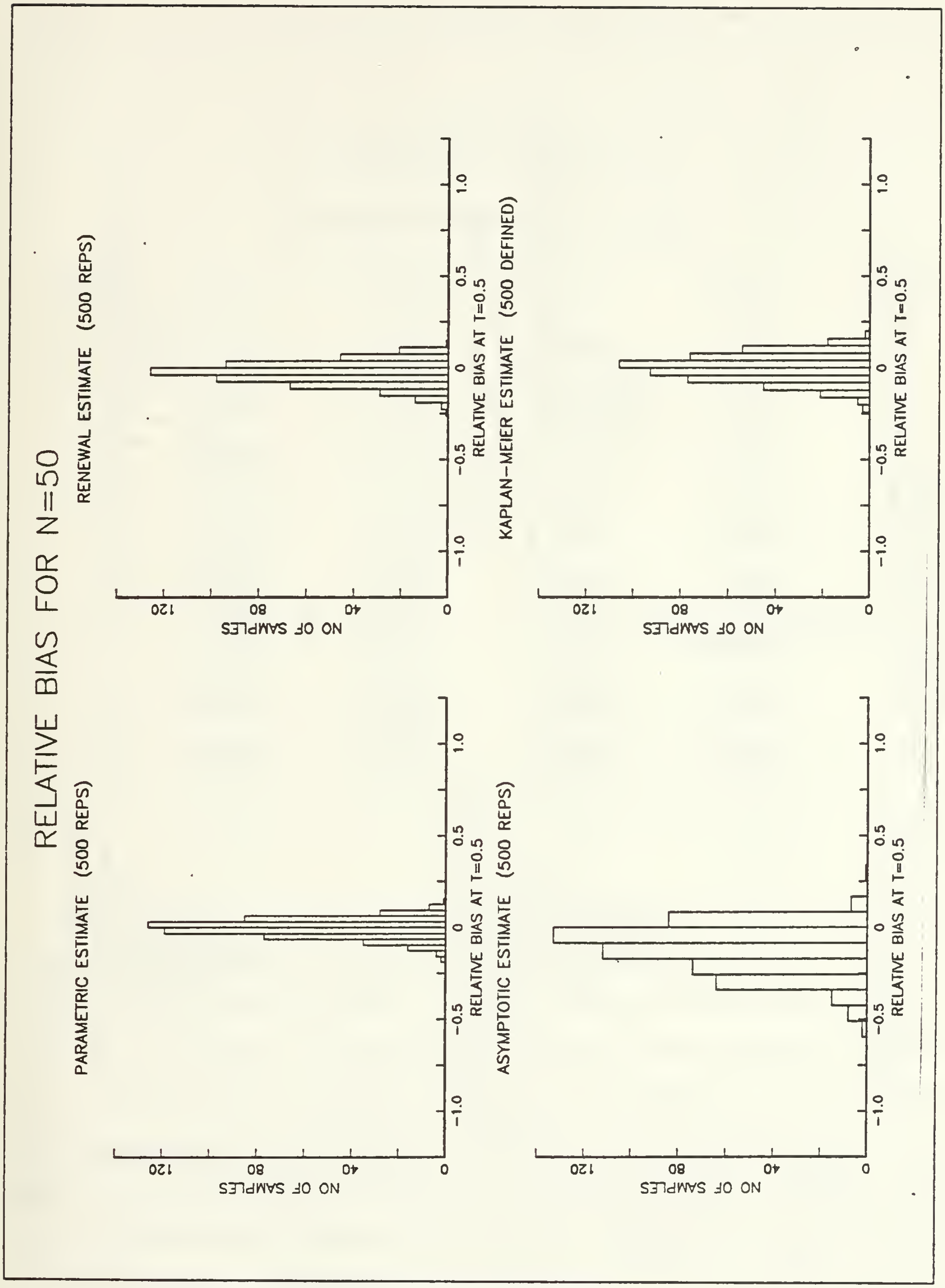

Figure 3.2a Histograms of relative bias for $N=50$ and $t=0.5$.

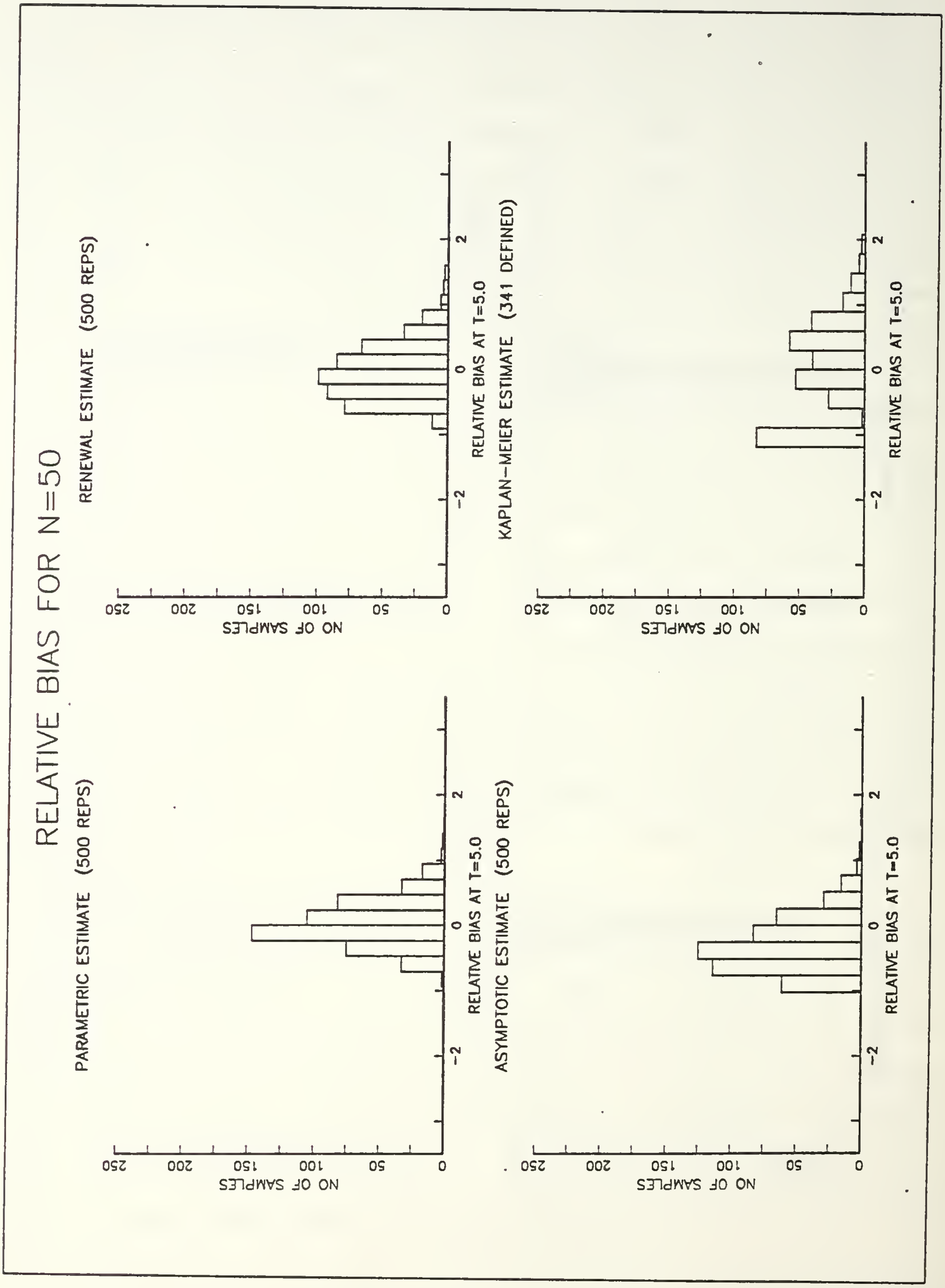

Figure 3.2b Histograms of relative bias for N= 50 and t= 5.0.

TABLE IV

AVERAGE RELATIVE BIAS

Exponential Model N = 50 (500 Reps)

Time	$\hat{P}_p(t)$	$\hat{P}_r(t)$	$\hat{P}_a(t)$	$\hat{P}_k(t)$	#KM
.5	-.00430 (.00398)	-.02928 (.00455)	-.12087 (.00743)	-.00082 (.00512)	500
1.0	-.00567 (.00220)	-.11500 (.00278)	-.12875 (.00580)	.00403 (.00330)	500
2.0	-.00364 (.00705)	-.28952 (.00742)	-.17511 (.01087)	-.00003 (.00872)	500
5.0	.03428 (.01623)	-.03348 (.01946)	-.29824 (.01876)	.01067 (.04022)	341
7.0	.08483 (.02326)	.56593 (.03139)	-.33861 (.02347)	-.35364 (.07749)	224
10.0	.19995 (.03633)	2.01703 (.07707)	-.35577 (.03175)	-.75760 (.08671)	188
12.5	.33775 (.05091)	4.36834 (.16959)	-.33896 (.04109)	-.72166 (.14855)	186
15.0	.52199 (.07067)	9.22171 (.36791)	-.29647 (.05409)	-1.00000 (.00000)	184

For N = 50, the parametric estimate again does well for small t. The average relative bias is within three standard deviations of zero for $t < 7$. An improvement for large t for the parametric estimate is expected because of the increased number of individuals. The renewal estimate again has large bias for large t, though not as much. The asymptotic estimate is negatively biased throughout time. For $t > 2$, the bias looks constant. For N = 50, the Kaplan-Meier estimate of the first passage time to state 0 has negligible average relative bias for $t < 7$. However as t increases, the Kaplan-Meier estimate looses an appreciable amount of its data due to undefined distributions. By $t = 15$, the Kaplan-Meier estimate has no survivors. Once again for large t ($t = 15$), the asymptotic estimate has the smallest average relative bias.

A simulation experiment is done for a case in which there is a relatively high number of individuals N = 100; $c=0.5$, $\rho_1=1$, $\rho_2=1$, $\theta=0.5$. The results appear in Table V. The increased number of individuals has decreased the average relative bias for all the estimates. The standard error of all the estimates has also decreased. From Tables III, IV, and V, it appears that as the number of observed individuals increase the average relative bias for all the estimates decrease.

TABLE V

AVERAGE RELATIVE BIAS

Exponential Model N = 100 (500 Reps)

Time	$\hat{P}_p(t)$	$\hat{P}_r(t)$	$\hat{P}_a(t)$	$\hat{P}_k(t)$	#KM
.5	-.00192 (.00150)	-.09574 (.00251)	-.09600 (.00399)	-.00379 (.00240)	500
1.0	-.00271 (.00275)	-.13534 (.00318)	-.08705 (.00512)	-.00235 (.00367)	500
2.0	-.00256 (.00495)	-.38676 (.00426)	-.10980 (.00756)	-.00267 (.00609)	500
5.0	.01258 (.01136)	-.69527 (.00808)	-.21245 (.01378)	.04573 (.02474)	448
7.0	.03523 (.01588)	-.51234 (.01557)	-.26003 (.01729)	-.15911 (.05919)	281
10.0	.08765 (.02337)	.28800 (.04112)	-.30374 (.02242)	-.47180 (.11283)	200
12.5	.14951 (.03059)	1.96444 (.09357)	-.32021 (.02708)	-.87080 (.09137)	185
15.0	.22930 (.03909)	5.69514 (.19744)	-.32195 (.03245)	-.87581 (.12419)	185

In order to investigate the effect of censoring on the values of $\hat{P}_r(t)$ and $\hat{P}_a(t)$ for large t, a simulation study is done in which the exponential censoring times has a mean of $1/c=1000$. The other parameters are $\rho_1=1$, $\rho_2=1$, and $\theta=0.5$ as before. Once again the number of individuals are N = 10, 50. The results are presented in Tables VI and VII. The results in Table VI suggest that the effect of the small sample size

resulting from N = 10 dominates the performance of all the estimates except for the Kaplan-Meier estimate for large t. The results of Table VII suggest that the method of computing $\hat{P}_r(t)$ is affecting its performance for large t since $\hat{P}_r(t)/\hat{P}_a(t) \sim 1$ as $t \rightarrow \infty$. For N = 50, limited censoring has improved the average relative bias of all the estimates for large t. Somewhat surprisingly, with limited censoring the Kaplan-Meier estimate has almost the best average relative bias. However, the standard error of its estimate is larger than that of the other estimates'. Thus the Kaplan-Meier estimate tends to be more variable than the other estimates.

TABLE VI

AVERAGE RELATIVE BIAS

Exponential Model N = 10 (Limited Censoring, c = 0.001) (500 Reps)

Time	$\hat{P}_p(t)$	$\hat{P}_r(t)$	$\hat{P}_a(t)$	$\hat{P}_k(t)$	#KM
.5	-.01895 (.00360)	-.01133 (.00441)	-.03853 (.00476)	-.00768 (.00718)	500
1.0	-.03008 (.00629)	-.04266 (.00644)	-.02281 (.00614)	-.00492 (.01025)	500
2.0	-.04178 (.01058)	-.08352 (.00955)	-.02283 (.01049)	-.01816 (.01461)	500
5.0	-.01755 (.02166)	-.02716 (.02091)	-.01934 (.02207)	.04571 (.02963)	500
7.0	.04003 (.02968)	.05500 (.03044)	.02483 (.03009)	.09415 (.03971)	500
10.0	.18364 (.04466)	.25792 (.04946)	.15076 (.04481)	.05896 (.06421)	500
12.5	.36140 (.06146)	.53159 (.07442)	.31332 (.06127)	.03902 (.09355)	500
15.0	.60294 (.08414)	.94267 (.11360)	.53732 (.08351)	-.00407 (.13365)	500

TABLE VII

AVERAGE RELATIVE BIAS

Exponential Model N = 50 (Limited Censoring, c = 0.001) (500 Reps)

Time	$\hat{P}_p(t)$	$\hat{P}_r(t)$	$\hat{P}_a(t)$	$\hat{P}_k(t)$	#KM
.5	-.00470 (.00144)	-.04348 (.00193)	-.07253 (.00267)	.00380 (.00313)	500
1.0	-.00760 (.00259)	-.15247 (.00285)	-.03784 (.00318)	.00104 (.00425)	500
2.0	-.01110 (.00452)	-.40891 (.00355)	-.01343 (.00464)	.00562 (.00645)	500
5.0	-.01035 (.00982)	-.79859 (.00278)	-.00883 (.00987)	.00497 (.01279)	500
7.0	-.00123 (.01340)	-.79946 (.00241)	-.00283 (.01350)	-.00350 (.01820)	500
10.0	.02459 (.01903)	-.59136 (.00335)	.01905 (.01922)	-.01116 (.02926)	500
12.5	.05757 (.02413)	-.13724 (.00679)	.04945 (.02440)	-.08114 (.04066)	500
15.0	.10115 (.02977)	.84405 (.01457)	.09097 (.03015)	-.06882 (.05631)	500

Below are reported simulation results experimenting with different parameter values of ρ_1 and c. For these studies, the number of individuals is set at $N = 50$ to reduce the effects of undefined Kaplan-Meier estimates of F_i. Four different cases are simulated. The sojourn time in state 1 is changed to reflect a higher and lower mean sojourn time and the censoring mean time is changed to reflect more or less censoring.

The first cases that are simulated are the changes in the mean sojourn time in state 1. The other parameters are $c = 0.5$, $\rho_2 = 1$, and $\theta = 0.5$ as before. The mean sojourn time of state 1 is increased from 1 to 2 ($\rho_1 = 0.5$) and decreased from 1 to 0.5 ($\rho_1 = 2$). With the increase in the mean sojourn time of state 1, the probability of a death being censored increases. For a decrease in the mean sojourn time, the opposite is true. There are quicker jumps out of state 1, resulting in more uncensored deaths. Tables VIII and IX show the computed average relative bias using equation 3.1 along with the associated standard error.

TABLE VIII

AVERAGE RELATIVE BIAS

Exponential Model N = 50 ($\rho_1 = 0.5$) (500 Reps)

Time	$\hat{P}_p(t)$	$\hat{P}_r(t)$	$\hat{P}_a(t)$	$\hat{P}_k(t)$	#KM
.5	.00045 (.00136)	-.00505 (.00188)	-.14754 (.00921)	.00062 (.00234)	500
1.0	.00175 (.00261)	-.02090 (.00326)	-.18477 (.01020)	.00707 (.00366)	500
2.0	.00647 (.00496)	-.06266 (.00597)	-.25529 (.01228)	.00877 (.00616)	500
5.0	.03886 (.01198)	.06045 (.01328)	-.40236 (.01702)	.07279 (.02202)	415
7.0	.07539 (.01717)	.25430 (.02093)	-.45897 (.01959)	-.04033 (.05350)	255
10.0	.15287 (.02643)	.76927 (.04165)	-.50727 (.02366)	-.65507 (.07690)	152
12.5	.24079 (.03615)	1.53886 (.07217)	-.52490 (.02772)	-.88639 (.06656)	137
15.0	.35370 (.04857)	2.79204 (.12186)	-.52738 (.03276)	-.86033 (.09855)	137

In Table VIII where the mean sojourn time in state 1 increases, the average relative bias of the parametric estimate looks about the same as in Table IV. The average relative bias of the renewal estimate is slightly better than in Table IV. The average relative bias of the asymptotic estimate looks like it increased, but is within three standard errors of Table IV. The average relative bias of the Kaplan-Meier estimate of the first passage time to state 0 looks the same as in Table IV. The number of defined Kaplan-Meier estimates has decreased due to the increase in the probability of a censored death as mentioned earlier. There are two survivors at $t = 15$.

In Table IX where the mean sojourn time in state 1 decreases, the parametric, asymptotic, and Kaplan-Meier estimates have the same average relative bias as in Table IV. The number of defined Kaplan-Meier estimates has increased due to the decrease in the probability of a censored death. The renewal estimate has increased as compared to Table IV.

TABLE IX

AVERAGE RELATIVE BIAS

Exponential Model $N=50$ $(\rho_1=2)$ (500 Reps)

Time	$\hat{P}_p(t)$	$\hat{P}_r(t)$	$\hat{P}_a(t)$	$\hat{P}_k(t)$	#KM
.5	-.00691 (.00323)	-.07878 (.00369)	-.11848 (.00504)	.00049 (.00434)	500
1.0	-.00907 (.00537)	-.22875 (.00512)	-.10384 (.00753)	-.00438 (.00685)	500
2.0	-.00826 (.00893)	-.49325 (.00721)	-.16288 (.01187)	-.00674 (.01211)	500
5.0	.04037 (.02060)	-.16543 (.02167)	-.30815 (.02132)	-.08353 (.05516)	308
7.0	.11149 (.03055)	.95936 (.04897)	-.34515 (.02811)	-.61487 (.08054)	221
10.0	.28446 (.05190)	5.05635 (.15416)	-.34050 (.04336)	-.78435 (.11361)	207
12.5	.50537 (.07987)	12.93165 (.40802)	-.28668 (.06490)	-1.00000 (.00000)	207
15.0	.82543 (.12395)	31.25238 (1.11064)	-.17898 (.10098)	-1.00000 (.00000)	207

The next cases that are simulated are the changes in the censoring distribution; the other parameters are $\rho_1=1$, $\rho_2=1$, and $\theta=0.5$ as before. The exponential mean time to censor is increased from 2 to 4 $(c=0.25)$ and decreased from 2 to 1 $(c=1)$. With an increase in the mean censoring time, the probability of a censored death decreases. With a decrease in the mean censoring time, the opposite is true. Tables X and XI show the average relative bias for each simulation along with the standard error.

Table X where the mean censoring time increases $(1/c=4)$, falls between Table IV and Table VII. The average relative bias of the parametric estimate is worse than in the limited censoring case of Table VII but slightly better than in the case $c=0.5$ of Table IV. The average relative bias of the renewal estimate is much worse than it is with the limited censoring of Table VII but about the same to slightly better in the tail

than in the case c=0.5 of Table IV. The average relative bias of the asymptotic estimate is about the same as the limited censoring case but much better than in the case c=0.5. The average relative bias of the Kaplan-Meier estimate of the first passage time to state 0 is about the same for small to moderate times and worse for large times than in the limited censoring case of Table VII and better for large times than in the case c=0.5 of Table IV. The number of defined Kaplan-Meier estimates is between the two tables, due to the increase in the mean censoring times. There are five survivors past $t = 15$.

TABLE X

AVERAGE RELATIVE BIAS

Exponential Model N=50 (c=0.25) (500 Reps)

Time	$\hat{P}_p(t)$	$\hat{P}_r(t)$	$\hat{P}_a(t)$	$\hat{P}_k(t)$	#KM
.5	-.00099 (.00179)	-.03349 (.00230)	-.06677 (.00331)	.00170 (.00325)	500
1.0	-.00071 (.00326)	-.13166 (.00348)	-.04538 (.00434)	.00874 (.00466)	500
2.0	.00178 (.00581)	-.36077 (.00525)	-.04544 (.00694)	.00837 (.00720)	500
5.0	.02794 (.01330)	-.57486 (.00892)	-.09177 (.01460)	.01752 (.02043)	486
7.0	.06206 (.01876)	-.27543 (.01723)	-.10565 (.01949)	.06881 (.04235)	411
10.0	.13879 (.02822)	.85619 (.03879)	-.09864 (.02727)	-.10377 (.09322)	307
12.5	.22881 (.03788)	2.71999 (.07635)	-.06942 (.03479)	-.59309 (.11491)	276
15.0	.34587 (.04987)	5.92070 (.15972)	-.01964 (.04386)	-.69629 (.16349)	273

In Table XI where the mean censoring time decreases, the average relative bias of the parametric estimate is about the same for small to moderate times and then is worse for large times than in Table IV. The average relative bias of the renewal and asymptotic estimates are both worse for $t > 2$ due possibly to an increase in the number of dishonest Kaplan-Meier estimates of F_i. The average relative bias of the the Kaplan-Meier estimate of the first passage time to state 0 is worse for $t > 5$. The number of defined Kaplan-Meier estimates has decreased reflecting the decrease in the mean time to censoring.

TABLE XI

AVERAGE RELATIVE BIAS

Exponential Model N = 50 (c = 1) (500 Reps)

Time	$\hat{P}_p(t)$	$\hat{P}_r(t)$	$\hat{P}_a(t)$	$\hat{P}_k(t)$	#KM
.5	-.00268 (.00270)	-.01350 (.00317)	-.27293 (.00971)	-.00012 (.00369)	500
1.0	-.00193 (.00491)	-.05220 (.00568)	-.35347 (.01171)	.00391 (.00610)	500
2.0	.00809 (.00878)	-.07041 (.01133)	-.47848 (.01452)	-.00203 (.01352)	488
5.0	.10233 (.02127)	.54674 (.03027)	-.66059 (.01795)	-.63062 (.07727)	128
7.0	.21600 (.03220)	1.46539 (.06044)	-.70866 (.02021)	-.93652 (.05174)	113
10.0	.47564 (.05588)	4.48047 (.15766)	-.73703 (.02540)	-1.00000 (.00000)	111
12.5	.80113 (.08693)	10.05152 (.34001)	-.73711 (.03215)	-1.00000 (.00000)	111
15.0	1.26789 (.13558)	21.62486 (.72038)	-.72108 (.04211)	-1.00000 (.00000)	111

C. ROBUSTNESS

In the above simulations the maximum likelihood estimate used the *known correct* model. Often, a model needs to be chosen to describe a data set. Attempts are made to analyze the data to determine a good model. However, when sample sizes are small, the difficulty of finding a good model increases. Hence due to small sample sizes or ease of computation, an incorrect model may be chosen to describe a data set. In this section, the robustness of the estimates proposed in Chapter II is studied with respect to an incorrect model assumption concerning the sojourn time in state 1.

The data for the simulation experiment in this section are generated from the following three state semi-Markov process: Individuals start in state 1 at $t=0$. The probability of a jump to state 0 is θ; to state 2 is $1-\theta$. From state 2 the probability of a jump to state 1 is 1. State 0 is an absorbing state. The sojourn time in state 2 is exponential with mean $1/\rho_2$. The sojourn time in state 1 is the sum of two independent exponentials with means $1/\rho_1$ and $1/\rho_3$; that is, the sojourn time in state 1 has a hypoexponential distribution. Censoring is independent and exponentially distributed with mean $1/c$. The same basic Fortran program is employed, modified for the above change. The data generated are analyzed by the same Fortran subroutines for each estimate as in the first section. In particular, the (incorrect) maximum likelihood estimate of equation 2.9 is used. This maximum likelihood estimate assumes the sojourn time in state 1 has an exponential distribution rather than the true hypoexponential distribution.

For the first simulation results reported, parameter values of $\rho_1=1$, $\rho_2=1$, $\rho_3=1$, $\theta=0.5$, and $c=0.5$ are used. Again, two different numbers of observed individuals are used, 10 and 50. The simulation is replicated 500 times and the average relative bias is computed utilizing equation 3.1. For the Kaplan-Meier estimate, M is taken as the number of defined Kaplan-Meier estimates of the first passage time to state 0 by time t. For the others, M is the number of replications. The actual value of the survivor function is computed by inverting the Laplace transform of the passage time to state 0 for the semi-Markov process.

Tables XII and XIII show the average relative bias of the hypoexponential model at selected values of t along with its associated standard error for N = 10 and 50. Again the right most column is the number of defined Kaplan-Meier estimates of the first passage time to state 0 out of 500 replications.

TABLE XII

AVERAGE RELATIVE BIAS

Hypoexponential Model N = 10 (500 Reps)

Time	$\hat{P}_p(t)$	$\hat{P}_r(t)$	$\hat{P}_a(t)$	$\hat{P}_k(t)$	#KM
.5	-.04965 (.00215)	.00603 (.00225)	-.18749 (.01511)	.00434 (.00332)	500
1.0	-.04224 (.00415)	.00518 (.00462)	-.27115 (.01478)	.00119 (.00622)	499
2.0	.02344 (.00829)	.01920 (.00994)	-.35821 (.01670)	.03759 (.01170)	488
5.0	.27429 (.02259)	.23906 (.02654)	-.45864 (.02448)	-.03008 (.05463)	253
7.0	.49966 (.03501)	.58490 (.04261)	-.46202 (.03062)	-.51308 (.08072)	164
10.0	.98106 (.06148)	1.53795 (.08272)	-.40812 (.04310)	-.90325 (.05049)	142
12.5	1.56115 (.09466)	2.93742 (.14238)	-.31726 (.05800)	-.97945 (.02055)	140
15.0	2.37288 (.14350)	5.27240 (.24293)	-.18077 (.07897)	-1.00000 (.00000)	139

In Table XII, for N = 10, the parametric estimate based on the incorrect model shows more relative bias for small t than the results in Table III using the correct maximum likelihood model. However, for moderate times $1 < t < 7$ the effect of the small number of individuals has overwhelmed the effect of the incorrect model and the relative bias is approximately the same as for the correct model given in Table III.

The average relative bias of the nonparametric estimates appear to do well. The renewal estimate and the Kaplan-Meier estimate of the first passage time to state 0 seem to do very well for small times and about the same for moderate to large times, with the Kaplan-Meier decreasing to -1.0 at $t = 15$. The asymptotic estimate seems to do about the same as in the situation of Table III; it is still negatively biased and has the smallest average relative bias for large times.

TABLE XIII

AVERAGE RELATIVE BIAS

Hypoexponential Model N = 50 (500 Reps)

Time	$\hat{P}_p(t)$	$\hat{P}_r(t)$	$\hat{P}_a(t)$	$\hat{P}_k(t)$	#KM
.5	-.04769 (.00104)	.00196 (.00104)	-.10608 (.00701)	-.00089 (.00154)	500
1.0	-.04010 (.00203)	-.00130 (.00229)	-.14888 (.00745)	.00136 (.00297)	500
2.0	.02140 (.00415)	-.02910 (.00520)	-.19231 (.00965)	.00557 (.00563)	500
5.0	.21701 (.01169)	.05244 (.01291)	-.29425 (.01620)	.03315 (.02328)	421
7.0	.36307 (.01803)	.20321 (.01998)	-.33710 (.01987)	-.09714 (.05655)	246
10.0	.63698 (.03053)	.64409 (.03839)	-.36628 (.02550)	-.53280 (.09457)	162
12.5	.92760 (.04476)	1.32451 (.06729)	-.36722 (.03096)	-.93675 (.05242)	150
15.0	1.29131 (.06391)	2.47259 (.11708)	-.35081 (.03768)	-.97947 (.02053)	149

In Table XIII, the case of the larger number of individuals N = 50, the effect of the incorrect model of the maximum likelihood estimate has a more noticable effect on the average relative bias; the average relative bias for the parametric estimate is significantly higher than for the nonparametric renewal and Kaplan-Meier estimates for $t \leq 2$. The nonparametric renewal and Kaplan-Meier estimates have about the same average relative bias for $t \leq 5$. The average relative bias of the asymptotic estimate does the same as in Table IV, and still consistently negatively biased. The Kaplan-Meier estimate of the first passage time to state 0 does the same as in Table IV. There is still one survivor at $t = 15$.

A simulation experiment is done for the case in which there is a relatively high number of subjects N = 100; the other parameters are as before. Table XIV show the effects of the increase in observed individuals. The average relative bias of the parametric estimate shows less relative bias than in Table XIII but still significantly higher than Table V using the correct model with comparable number of subjects. The average relative bias of the nonparametric estimates has lower relative bias than Table XIII and about the same relative bias as Table V. Again it appears that the average relative bias for all the estimates decrease as the number of individuals increase.

TABLE XIV

AVERAGE RELATIVE BIAS

Hypoexponential Model N = 100 (500 Reps)

Time	$\hat{P}_p(t)$	$\hat{P}_r(t)$	$\hat{P}_a(t)$	$\hat{P}_k(t)$	#KM
.5	.00010 (.02439)	.04879 (.02528)	-.04324 (.01982)	.05034 (.02493)	500
1.0	.03696 (.03963)	.05200 (.03902)	-.05118 (.03019)	.07747 (.03903)	500
2.0	.13258 (.05843)	-.05690 (.04927)	-.05490 (.03941)	.11607 (.05859)	500
5.0	.18619 (.00875)	-.44921 (.01064)	-.21495 (.01202)	-.02024 (.01586)	491
7.0	.30873 (.01312)	-.18768 (.01709)	-.26080 (.01481)	-.04054 (.03720)	335
10.0	.52842 (.02135)	.50623 (.02773)	-.30765 (.01864)	-.47076 (.08035)	184
12.5	.75156 (.03012)	1.15119 (.04520)	-.33121 (.02174)	-.85091 (.06960)	160
15.0	1.01691 (.04147)	2.06055 (.07748)	-.34404 (.02499)	-.92400 (.05621)	158

To study the effects of censoring for this model, a simulation experiment is done in which the exponential censoring times has a mean of $1/c = 1000$. The other parameters are $\rho_1 = 1$, $\rho_2 = 1$, $\rho_3 = 1$, and $\theta = 0.5$ as before. The number of individuals is N = 50. The results are shown in Table XV. The average relative bias of the parametric

estimate is higher for all times than Table VII using the correct model. The average relative bias of the nonparametric estimates are about the same as Table VII. Again, even with limited censoring, the average relative bias of the renewal estimate has computational problems for large t.

TABLE XV

AVERAGE RELATIVE BIAS

Hypoexponential Model N=50 (Limited Censoring, c=0.001) (500 Reps)

Time	$\hat{P}_p(t)$	$\hat{P}_r(t)$	$\hat{P}_a(t)$	$\hat{P}_k(t)$	#KM
.5	-.07105 (.00060)	-.00384 (.00081)	-.03344 (.00157)	.00197 (.00136)	500
1.0	-.08365 (.00114)	-.03442 (.00156)	-.04223 (.00170)	.00167 (.00226)	500
2.0	-.06224 (.00224)	-.19090 (.00279)	-.02480 (.00254)	.00162 (.00391)	500
5.0	-.00469 (.00561)	-.70151 (.00294)	-.00948 (.00610)	.00063 (.00789)	500
7.0	.02631 (.00798)	-.80405 (.00221)	-.01056 (.00856)	-.00350 (.01105)	500
10.0	.07973 (.01182)	-.75567 (.00230)	-.00598 (.01231)	.00391 (.01615)	500
12.5	.13071 (.01535)	-.59906 (.00345)	.00328 (.01553)	-.01794 (.02164)	500
15.0	.18831 (.01925)	-.31213 (.00598)	.01761 (.01889)	-.02465 (.02797)	500

Two additional simulations are done using different hypoexponential distributions for the sojourn time in state 1. For these simulations the number of individuals is set at N=50 for comparative purposes. The first simulation uses a hypoexponential distribution of $\rho_1=1$, $\rho_2=1$, $\rho_3=0.1$, $\theta=0.5$, and $c=0.5$. Table XVI shows the average relative bias and standard error for this model.

TABLE XVI

AVERAGE RELATIVE BIAS

Hypoexponential Model N = 50 ($\rho_3 = 0.1$) (500 Reps)

Time	$\hat{P}_p(t)$	$\hat{P}_r(t)$	$\hat{P}_a(t)$	$\hat{P}_k(t)$	#KM
.5	-.01070 (.00036)	.00048 (.00033)	-.41461 (.01923)	-.00004 (.00049)	500
1.0	-.01378 (.00071)	-.00012 (.00079)	-.49394 (.01835)	-.00087 (.00108)	500
2.0	-.00851 (.00141)	-.00275 (.00172)	-.59801 (.01730)	-.00309 (.00208)	500
5.0	.03290 (.00359)	-.00058 (.00541)	-.74519 (.01537)	-.00397 (.00716)	492
7.0	.06534 (.00512)	.01743 (.00817)	-.79285 (.01446)	.01353 (.01382)	372
10.0	.11795 (.00755)	.08352 (.01168)	-.83604 (.01349)	-.08234 (.04149)	165
12.5	.16554 (.00971)	.17542 (.01464)	-.85869 (.01295)	-.39425 (.07374)	85
15.0	.21680 (.01203)	.28205 (.01784)	-.87462 (.01258)	-.69274 (.08075)	60

Surprisingly, in Table XVI, the average relative bias of the parametric estimate is significantly better than Table XIII. The survivor function of the first passage time to state 0 for the semi-Markov model having the sum of two exponentials with mean 1 and 10 for the sojourn time in state 1 was computed. It was compared to the corresponding survivor function of the Markov model of Chapter II having exponential sojourn time in state 1 with mean 11. The parameters $\rho_2 = 1$, and $\theta = 0.5$ are as before for both models. For large t, the two survivor functions are approximately $P\{D > t\} = \exp[-0.045t]$ for the semi-Markov model and $P\{D > t\} = \exp[-0.043t]$ for the Markov model. Thus it appears that the small average relative bias for the parametric estimate in Table XVI is due to the closeness of the survivor functions for the two models. The average relative bias of the renewal estimate is significantly better than Tables XIII and IV. The average relative bias of the asymptotic estimate is significantly worse than Tables XIII and IV. A possible explanation for this is that with a mean sojourn time in state 1 of approximately 11 and a mean censoring time of 2, the process either jumps to state 0 at first transition or

becomes censored due to the expected long sojourn time in state 1. Therefore, the Kaplan-Meier estimates for F_i will probably contain most of the probability mass at small times and relatively little mass for large times due to censoring, causing the Kaplan-Meier estimates of F_i to be unreliable for large times. The renewal and Kaplan-Meier estimates have approximately the same relative bias for $t \leq 7$. The average relative bias of the Kaplan-Meier estimate of the first passage time to state 0 is about the same as Tables XIII and IV. For $t \geq 12.5$, the Kaplan-Meier estimate is significantly better than Tables XIII and IV, however, of the 500 replications only 85 are still defined by $t = 12.5$ in Table XVI.

The next simulation experiment uses a hypoexponential distribution for the sojourn time in state 1 that very closely resembles the exponential distribution used in Table IV. For this simulation, the parameter are $\rho_1 = 1$, $\rho_2 = 1$, $\rho_3 = 100$, $\theta = 0.5$, and $c = 0.5$. Again $N = 50$ for comparison purposes. Table XVII show the average relative bias and standard error of the estimates at selected times. The average relative bias of all of the estimates are about the same as in Table IV as expected.

TABLE XVII

AVERAGE RELATIVE BIAS

Hypoexponential Model $N = 50$ $(\rho_3 = 100)$ (500 Reps)

Time	$\hat{P}_p(t)$	$\hat{P}_r(t)$	$\hat{P}_a(t)$	$\hat{P}_k(t)$	#KM
.5	-.00825 (.00213)	-.02784 (.00266)	-.12817 (.00574)	-.00435 (.00334)	500
1.0	-.01009 (.00387)	-.11168 (.00420)	-.13732 (.00743)	-.00655 (.00473)	500
2.0	-.01009 (.00684)	-.30050 (.00717)	-.18659 (.01074)	-.00933 (.00880)	500
5.0	.02098 (.01543)	-.05434 (.01872)	-.31762 (.01803)	.02698 (.03958)	333
7.0	.06513 (.02182)	.52736 (.03064)	-.36331 (.02223)	-.18712 (.08288)	213
10.0	.16666 (.03337)	1.90635 (.07764)	-.38933 (.02940)	-.61354 (.11659)	166
12.5	.28695 (.04577)	4.30938 (.16887)	-.38193 (.03724)	-.79971 (.14178)	161
15.0	.44578 (.06199)	8.99799 (.36461)	-.35141 (.04795)	-1.00000 (.00000)	160

IV. CONCLUSIONS

From the results of Chapter III, it can be concluded:

1) The maximum likelihood estimate uses the most assumptions about the model. It understandably does well when the model used is correct. It is the most sensitive to incorrect model assumptions.

2) The renewal estimate and asymptotic estimate are biased by censoring of the last sojourn time in a state which makes the Kaplan-Meier estimate undefined. Further analysis could be done to investigate reasonable methods to make the Kaplan-Meier estimate honest.

3) The asymptotic estimate has the smallest average relative bias for large times, $t = 15$. However, the bias is always negative. Further analysis could be done to find a bias correction for it.

4) The Kaplan-Meier estimate of the first passage time to state 0 uses the least knowledge. It does well for small times and moderate to large numbers of individuals. The Kaplan-Meier estimate and the renewal estimate appear to do about as well for small t.

5) The larger the number of individuals the smaller the average relative bias is for all the estimates.

6) The renewal estimate requires a great deal of computation. In view of the simulation results, one recommendation is to use the Kaplan-Meier estimate as long as not too many observations are censored, and then use the asymptotic estimate for larger times. The asymptotic estimate needs to be used with caution if the last sojourn times in state 1 or 2 are censored.

LIST OF REFERENCES

1. Ross, S.M. (1985). *Introduction to Probability Models.* 3rd ed. Academic Press, Inc. Orlando, Florida.

2. Cox, D.R. (1984). Some remarks on semi-Markov processes in medical statistics. *International Symposium on Semi-Markov Processes and their Applications.* Brussels, Belgium.

3. Weiss, G.H. and Zelen, M. (1965). A semi-Markov model for clinical trials. *Journal of Applied Probability*, **2**, 269-285.

4. Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481.

5. Jacobs, P. A. (1986). Results of a simulation study of estimates of a first passage time distribution for censored semi-Markov process (unpublished).

6. Lewis, P.A.W. and Uribe, L. (1981). The New Naval Postgraduate School Random Number Package - LLRANDOM II. Naval Postgraduate School Technical Report, NPS 55-81-005. Monterey, California.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Technical Information Center Cameron Station Alexandria, Virginia 22304	2
2.	Library, Code 0142 Naval Postgraduate School Monterey, California 93943	2
3.	Prof. P.A. Jacobs, Code 55Jc Department of Operations Research Naval Postgraduate School Monterey, California 93943	1
4.	Prof. J.D. Esary, Code 55Ey Department of Operations Research Naval Postgraduate School Monterey, California 93943	1
5.	Prof. D.P. Gaver, Code 55Gv Department of Operations Research Naval Postgraduate School Monterey, California 93943	1
6.	Lt. R.M. Gallagher 6827 18th Av. S. Richfield, Minnesota 55423	2

220219

Thesis
G14045 Gallagher
c.1 A simulation study of estimates of a first passage time distribution for a censored semi-Markov process.